

Entrepôt de données

Par Maggie LEKPA

INTRODUCTION

Depuis les années 80, les entreprises ont vu la quantité de données évoluée à la vitesse grand V pour atteindre en 2010 → 2 milliards de téraoctets

En 2015 → x5

Entre 2020 et 2025 → x3,7

2035 → 2142 zettaoctets

Très vite → tirer profit des données; les analyser pour en tirer profit notamment pour savoir ce qui marche ou ne marche pas bien; ce qui est à améliorer.

INTRODUCTION

Besoin d'un système d'aide à la décision :

Simple et Rapide : les personnes chargées de prendre les décisions ne sont pas forcément des informaticiens et ont un très précieux.

Performant : afin de traiter de gros volume de données

Indépendant du système de production : important de dissocier la production avec le système décisionnel

Gestion d'accès: tout le monde n'est pas habilité à prendre les décisions dans l'entreprise.

Hétérogène : toutes les données de l'entreprise doivent pouvoir y être agrégées pour une meilleure prise de décision.

INTRODUCTION

Les systèmes décisionnels sont nés...

Deux mondes appelés à coexister :

Le monde **opérationnel** et le monde **décisionnel**

Système OPERATIONNEL (OLTP)

OLTP : OnLine Transactionnel Processing

Destiné à l'ensemble des opérations (transactions) de l'entreprise (traitement d'une commande, gestion des stocks, édition d'une facture...)

Faible volumétrie de donnée : donnée de l'entreprise à l'instant présent.

Est en lecture, écriture et modification

Système DECISIONNEL

Repose sur une architecture BI (Business Intelligence)

Aide à piloter l'entreprise et à prendre des décisions: tendance, moyennes, écart type des indicateurs de l'entreprise...

Gros volume de données : données historique sur plusieurs mois, années..

En lecture seule

Intégration de données

Une entreprise peut avoir plusieurs systèmes opérationnels :
CRM (Client Relationship Management) pour la gestion des clients

ERP (Entreprise Ressource Planning) pour la gestion des ressources...

Système de gestion des stocks/commandes...

Besoin de mettre toutes les données en relation (dans un lieu unique) pour pouvoir l'analyser et en tirer la valeur.

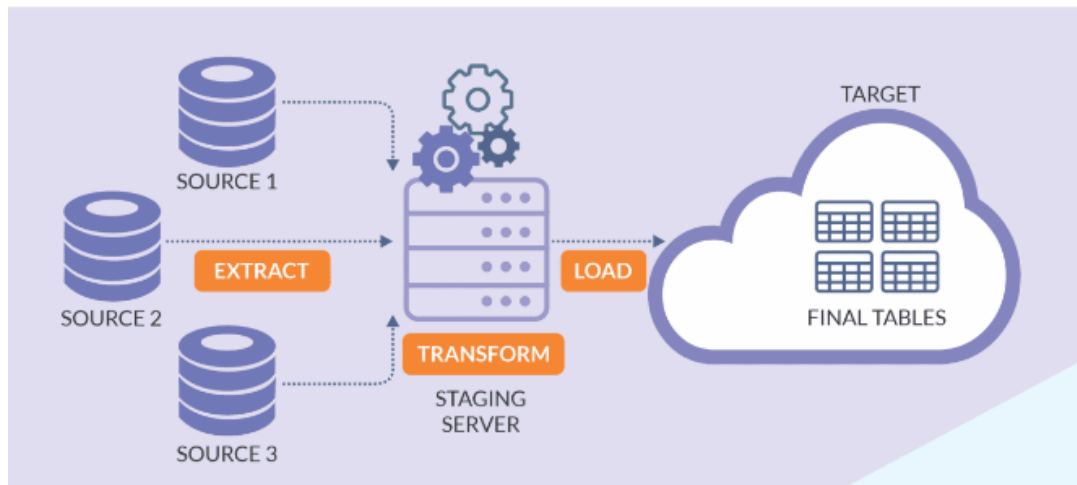
Intégration de données

Comment quitte t'on de un ou plusieurs systèmes opérationnels à un système unique décisionnel? → **ETL (Extract Transform Load)**

ETL est un système par lequel les données vont transiter du système opérationnel au système décisionnel. Les données vont être nettoyées et ou transformées avant d'être stockées.

Intégration de données - ETL

ETL (Extract Transform Load)

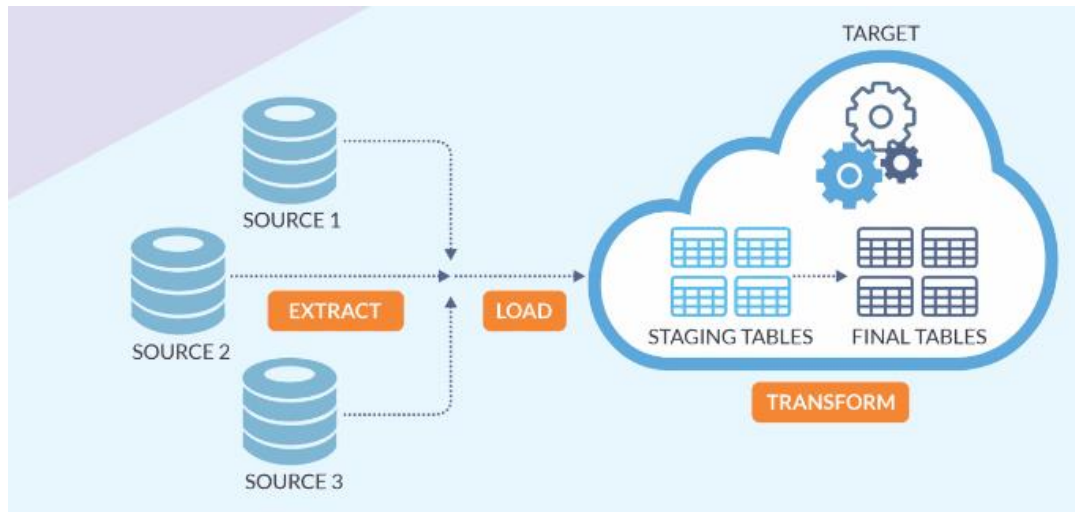


Système opérationnel

Système décisionnel

Intégration de données

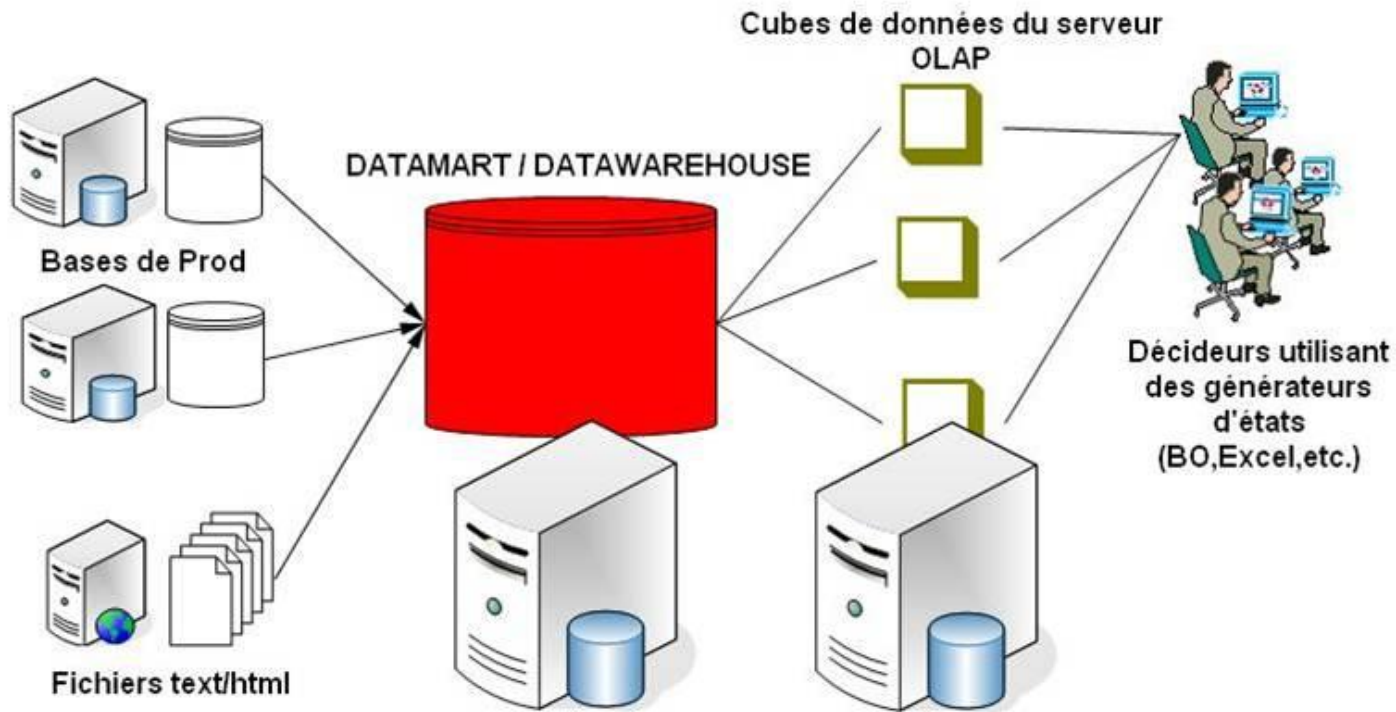
Variante : ELT (Extract Load Transform)



Intégration de données

ETL	ELT
La transformation est faite avant le chargement dans les tables → données sensibles masquées ou supprimées	Transformation faite après le chargement dans les tables → demande un bon niveau de sécurité
Le chargement des données prend du temps	Données disponibles très rapidement.
Gestion de l'évolution des besoins de données difficile à gérer.	Plus flexible car toutes les données sont à disposition

Architecture BI



Architecture BI

Architecture BI

Une archi BI est un ensemble d'éléments qui mis ensemble permettent de créer de la connaissance (de la valeur).

Elle est constitué d'un **entrepôt de données** qui contient toutes les données de l'entreprise; données nettoyées et transformées.

Organisé en axe d'analyse (**table de dimension**) et objets d'analyse (**table de fait**) avec différent niveau de granularité.

Entrepôt de données est un **système intégré** →

Architecture BI – entrepôt de données

Système intégré → les données stockées dans l'entrepôt proviennent des différentes sources de données de l'entreprise.

Ces sources pouvant être des feuilles de calculs, des fichiers plats, base Oracle/MySQL ...

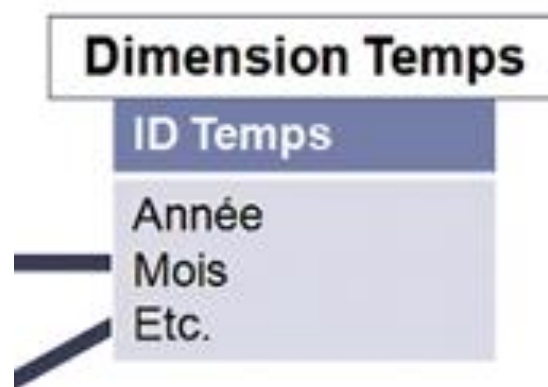
L'intégration se fait via des ETL (**E**xtract **T**ransform **L**oad) ou ELT

Architecture BI

Base de données classique → relation (table)

Architecture BI → dimension et table de fait

Dimension : ce qu'on utilise pour faire l'analyse (le temps, par société, par poste...)



Architecture BI

Base de données classique → relation (table)

Architecture BI → dimension et table de fait

Table de fait : contient les données à analyser; ce sont des valeurs, quantité (le chiffre d'affaire, la rémunération...)



Architecture BI

Fait : ce qu'on analyse; c'est une valeur quantitative
(le chiffre d'affaire, la rémunération...)

Dimension : ce qu'on utilise pour faire l'analyse (le temps, par société, par poste...)

Les tables de fait et de dimension peuvent être mises en relation de deux façons différentes : **Etoile** et **flocon**.

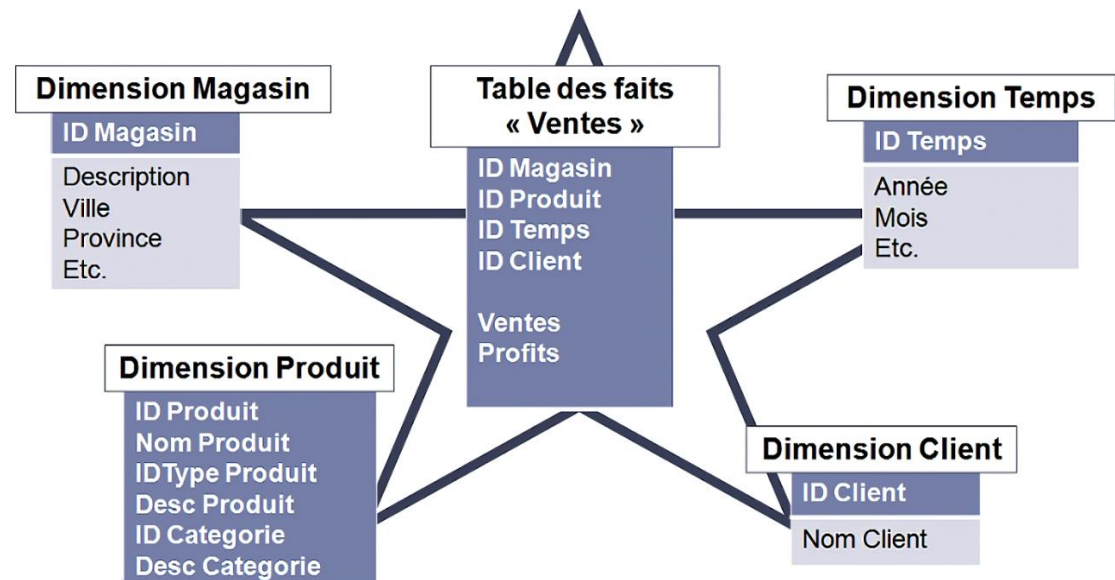
Architecture BI – Modélisation en étoile

Modélisation en **Etoile** :

Une table de fait centrale

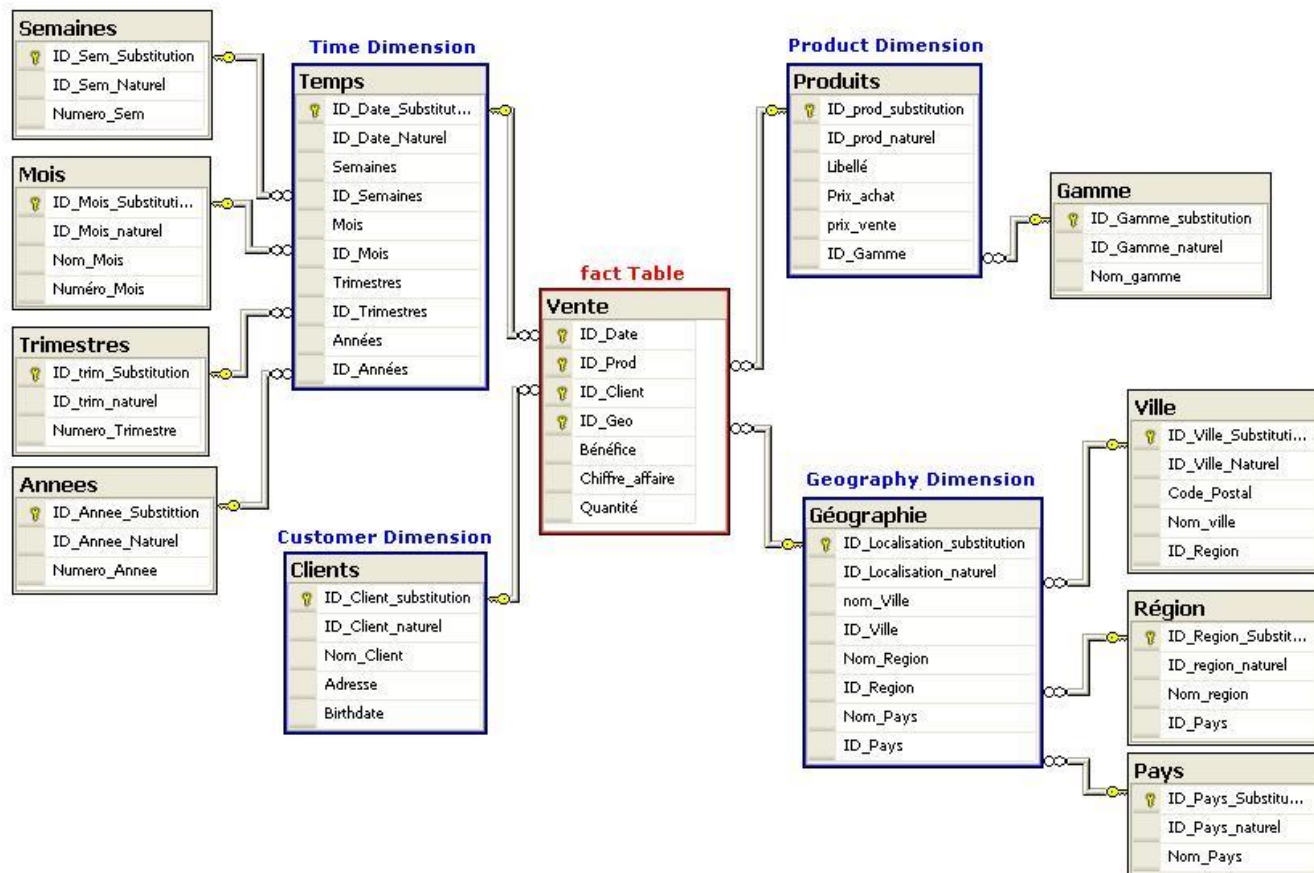
Chaque dimension est directement relié à un fait

Pas de liaison entre deux dimension



Architecture BI – Modélisation en flocon

Modélisation en **Flocon** : certaines dimensions peuvent être reliées à d'autres dimensions créant ainsi une hiérarchie.



Modélisation d'un entrepôt de données

Un entrepôt de données est un ensemble de tables de faits et de dimensions modélisé en étoile et ou flocon

Entrepôt de données : est une constellation d'étoiles et ou de flocon reliés entre eux via des dimensions.

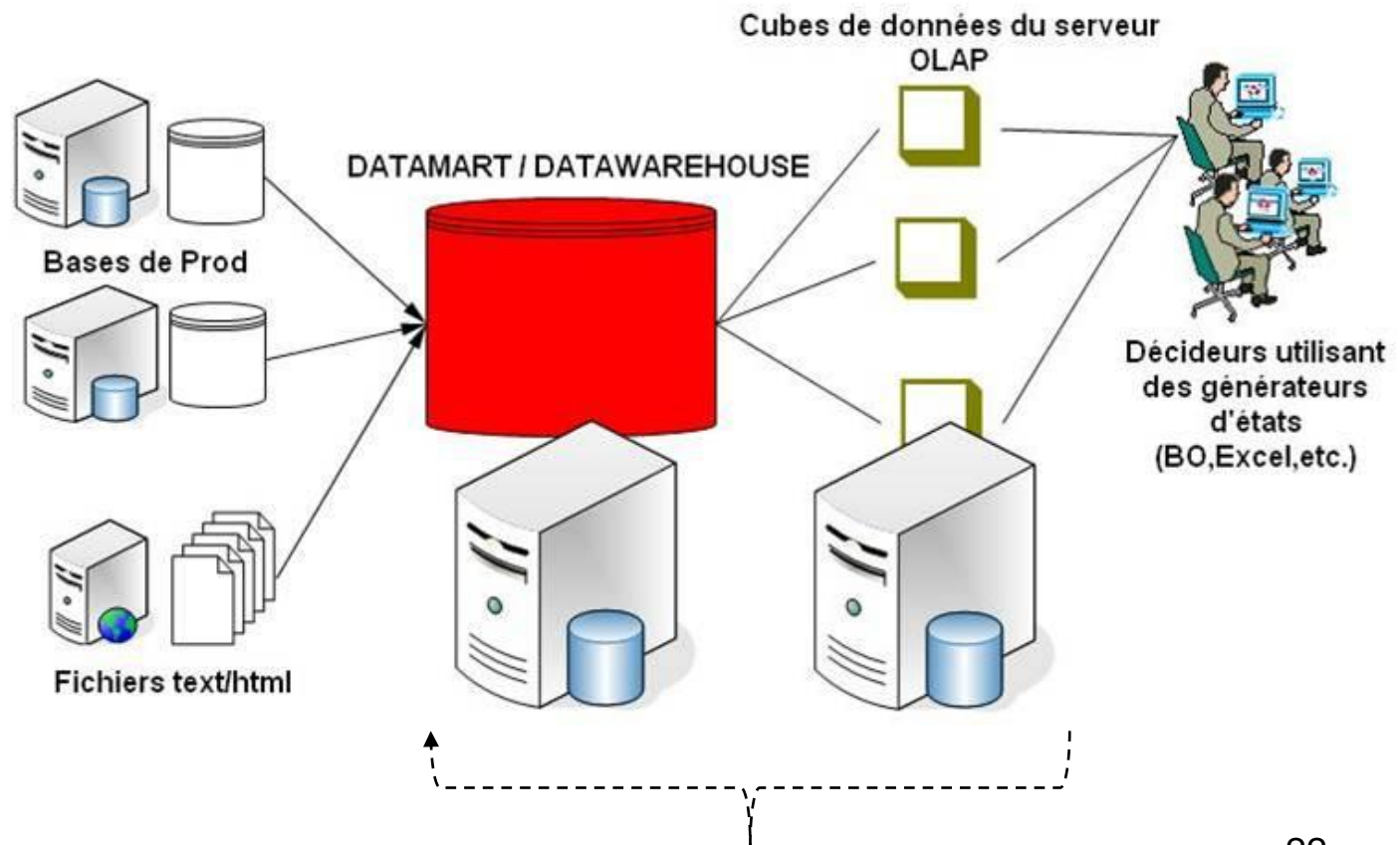
Comment les étoiles sont elles regroupées?

Modélisation d'un entrepôt de données

Trois approches:

- **TOP-DOWN** : on conçoit toutes les étoiles qui constitue l'entrepôt puis on le modélise
- **BOTTOM-UP** : les étoiles sont créées au fur et à mesure puis regroupées
- **MIDDLE-OUT** : approche hybride; conception de la totalité de l'entrepôt, division en petite entité, mise en place de chaque entité

Architecture BI - DATAMART



Architecture BI - DATAMART

Les entrepôts de données étant très volumineux → besoin de les diviser en élément plus petits et facile à gérer → **DATAMART**

Un DATAMART est un sous ensemble d'un entrepôt; il contient les données liées à une fonction de l'entreprise (vente, client..) ou a une branche de l'entreprise.

Moins volumineux

Architecture BI – cube OLAP

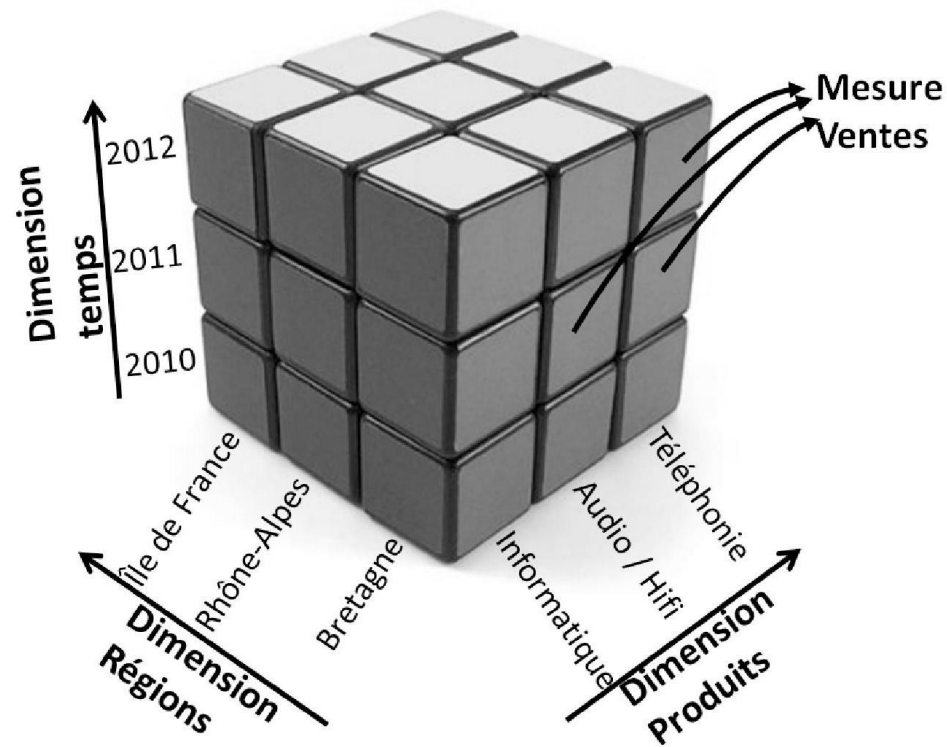
Les datamarts utilisent généralement les technologies OLAP (OnLine Analytical Processing) pour structurer les données.

OLAP est une base de données multidimensionnelle (cube ou hypercube).

Il est construit pour permettre aux analystes d'explorer les données de l'entrepôt suivant différents axes; de faire une multitude de croisements.

Exemple : calcul du chiffre d'affaire par client, par produit et par zone géographique.

Architecture BI – cube OLAP



Architecture BI – cube OLAP

Différents types de systèmes

MOLAP : Multidimensional OnLine Analytical Processing

Les données sont précalculées et stockées dans des cubes → accès rapide

Quantité de données limitées car pas possible d'inclure un grand volume de données dans le cube

ROLAP : Relationnal OLAP

la vue multidimensionnelle est créée de façon dynamique via des requêtes sql → accès plus lent

Gestion de gros volumes de données

Architecture BI – cube OLAP

Différents types de systèmes

HOLAP : Hybrid OLAP

Mélange de ROLAP et MOLAP

ROLAP pour accéder rapidement aux données agrégées

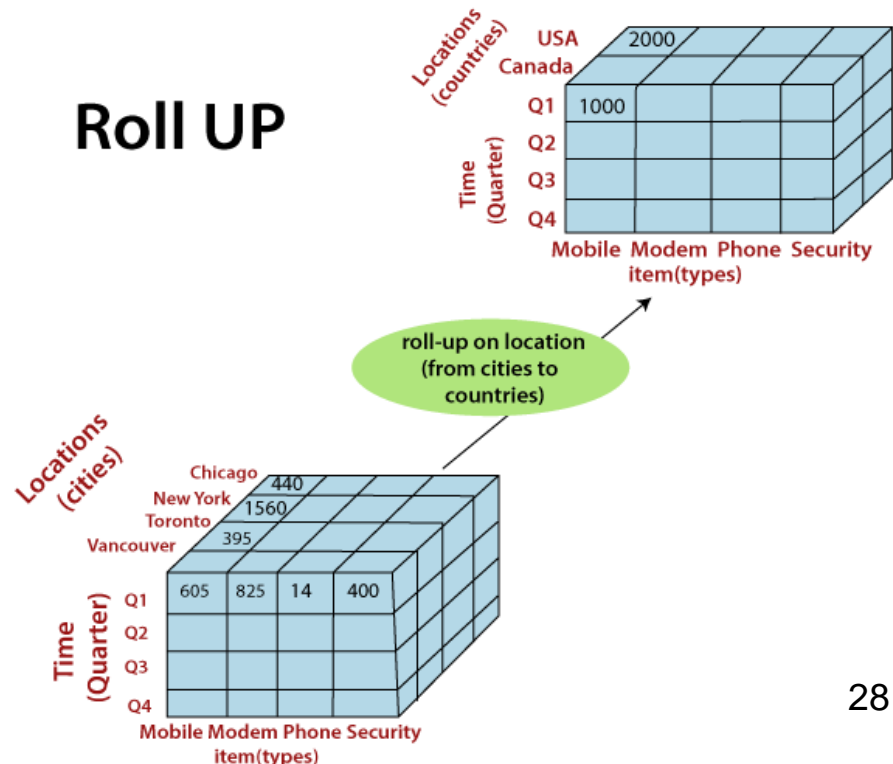
MOLAP pour obtenir des détails très fins

Architecture BI – cube OLAP

Plusieurs techniques d'analyse :

Roll-up (drill up) : opération qui permet de rassembler les données

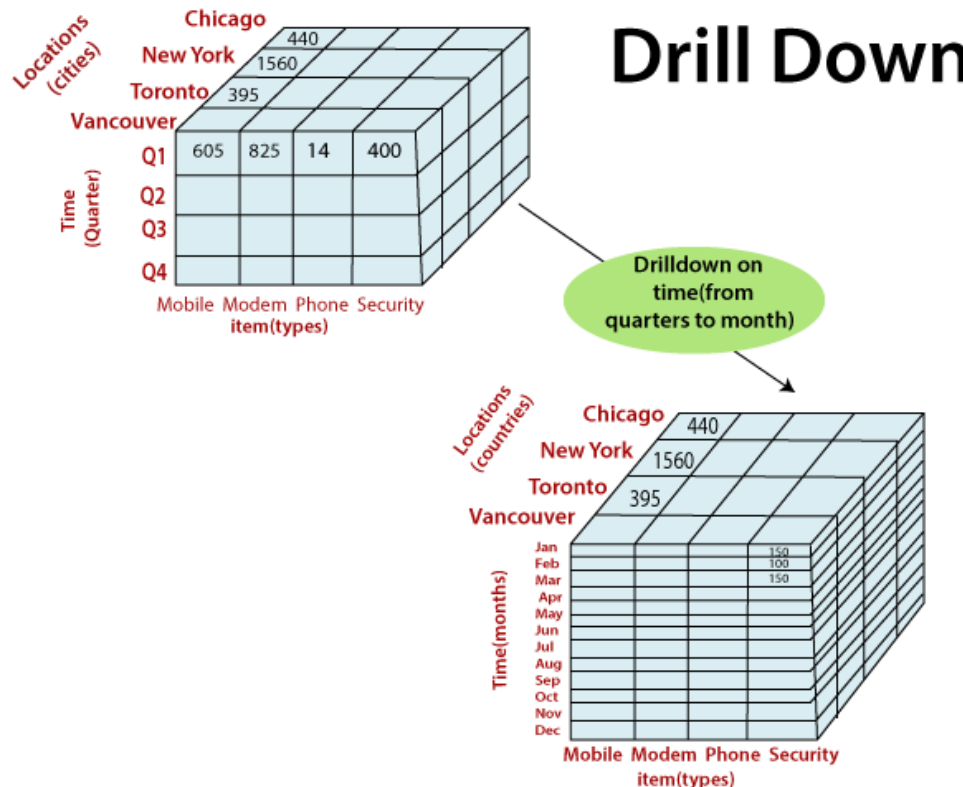
Roll UP



Architecture BI – cube OLAP

Plusieurs techniques d'analyse :

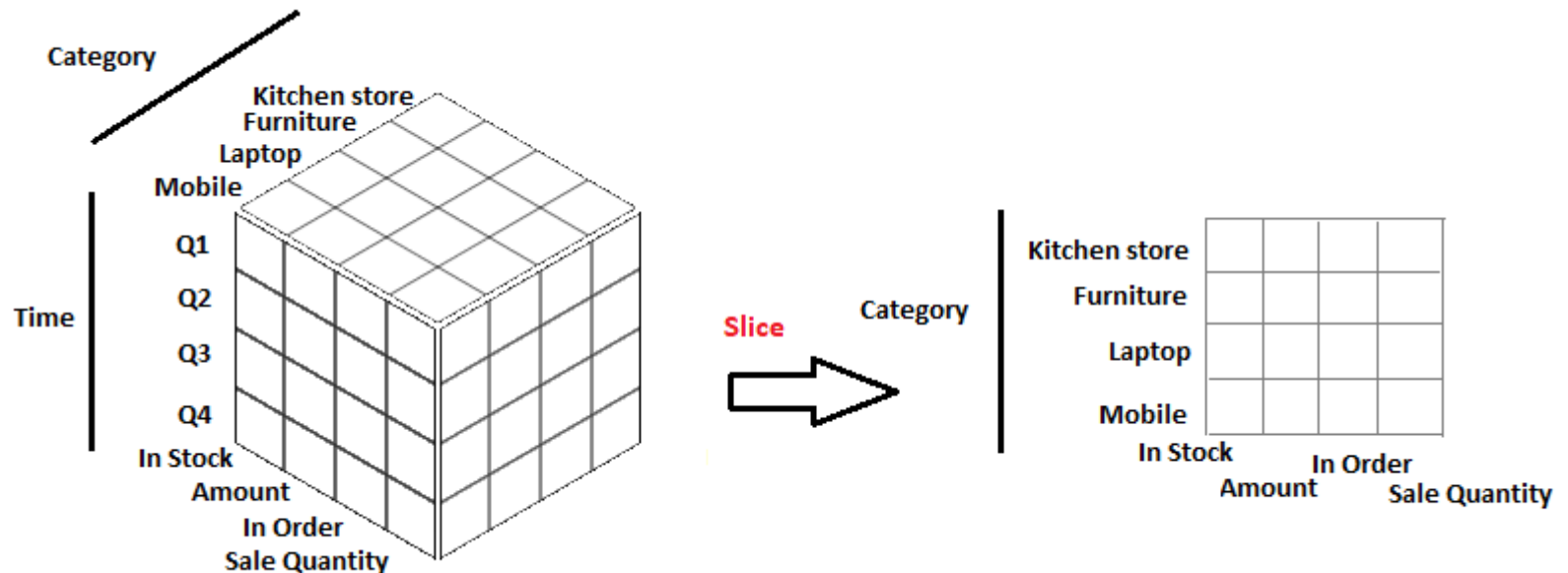
Drill-down: opération qui permet d'affiner l'analyse.



Architecture BI – cube OLAP

Plusieurs techniques d'analyse :

SLICE: opération qui consiste pour une valeur d'un axe, d'obtenir le sous cube (dimension 2)



Outil d'analyse et de restitution

Un fois l'entrepôt de données mis en place et les datamarts créés
-→ visualiser les données pour prise de décisions.

Outils de visualisation : PowerBI

Fin